Neural Processing Letters (2022) 54:5333–5349
https://doi.org/10.1007/s11063-022-10864-z

Traffic Scene Perception Based on Joint Object Detection and Semantic Segmentation

Libo Weng[1] · Yingjie Wang[1] · Fei Gao[1]

Accepted: 22 April 2022 / Published online: 4 June 2022

Abstract
Traffic scene visual perception technology is very important for intelligent transportation. Although the emerging panoptic segmentation is the most desirable sensing technology, object detection and semantic segmentation are relatively more mature and have fewer requirements for data annotation. In this paper, a joint object detection and semantic segmentation perception method is proposed for both practicability and accuracy. The proposed method is based on the results of object detection and semantic segmentation. Firstly, the result of basic semantic segmentation is preprocessed according to the principle of entropy. Secondly, the candidate bounding boxes of pedestrians and vehicles are extracted by object detection. Thirdly, candidate bounding boxes are optimized by using a K-means based vertex clustering algorithm. Finally, the contours of scene elements are matched with the results of semantic segmentation. The experimental results on the Cityscapes dataset show that the final perception effect is more susceptible to semantic segmentation results. The theoretical upper limit of the actual perception effect is 95.4% of the ground-truth of panoptic segmentation. The proposed method can effectively combine object detection and semantic segmentation, and achieve perception results similar to panoptic segmentation without additional data annotation.

Keywords Object detection · Semantic segmentation · Joint perception · Panoptic segmentation

1 Introduction

Visual perception of traffic scenes is one of the research hotspots in the field of intelligent transportation. Since there are a large number of elements in the actual traffic scene, it is a huge challenge to achieve holistic scene perception. Referring to an early study [1], traffic scene elements can be divided into two categories: things class, i.e., countable elements (e.g., pedestrians, vehicles and animals) and stuff class, i.e., the same texture area (e.g., sky, road

✉ Fei Gao
feig@zjut.edu.cn

[1] College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310014, China

and grass). After decades of research, a lot of great progress has been made in the visual perception of traffic scenes. Before the ubiquitous use of deep learning, researchers targeted specific perception tasks, such as the detection and recognition of vehicles, pedestrians and traffic signs. Although holistic traffic scene perception has profound practical significance, it did not become a research hotspot due to the limitation of knowledge and technology at that time. The development of deep learning promotes the perception effect to a new level. In addition to single-task oriented deep neural networks, multi-task oriented deep neural networks are spring up continually, showing amazing scene perception performance. Panoptic segmentation [2] woke up the research interests in holistic traffic scene perception.

For intelligent transportation, it is difficult to achieve both practicability and high accuracy of visual perception algorithms. At present, in the field of visual perception of traffic scenes, object detection [3–7] and semantic segmentation [8–15] algorithms are relatively mature and useful, whose requirements for data annotation are also simple. However, object detection algorithms often produce false positive and false negative results, and cannot obtain the contours of the objects. While semantic segmentation algorithms can obtain the contours of the scene elements, but may have wrong results in image pixel classification and cannot identify the individuals in the object groups (individual person in the crowd, separated car in the vehicle group, etc.). Instance segmentation algorithms [16] can achieve the identification of the individual objects in the groups by combining object detection and semantic segmentation, but only focus on the foreground objects in the scene while ignoring the background elements. Panoptic segmentation algorithms [2, 17–28], which can compensate for the disadvantages of instance segmentation, achieve the perception of the whole scene by combining semantic segmentation and instance segmentation. However, panoptic segmentation algorithms require complex data annotation detailed to the contour of each object in the scene. Due to the complexity and variability of the actual traffic scene, the workload of data labeling is huge. The number of labeled samples for model training in deep learning is relatively large, even though training samples can be generated by generative adversarial networks which are different from the real samples captured by cameras. Panoptic segmentation algorithms have a promising future, but still lack practicality currently.

In this paper, a joint object detection and semantic segmentation algorithm is proposed. The practicability of the method is guaranteed, since the data labeling only for basic object detection and semantic segmentation is needed, which is quite simpler than that for panoptic segmentation. At the same time, similar accuracies of the perception results to panoptic segmentation can be achieved. Firstly, under the principle of information entropy, semantic binary images of pedestrians and vehicles can be extracted from the basic semantic segmentation results, and then the processing of denoising and enhancing will be carried out. Secondly, the candidate bounding boxes of pedestrians and vehicles will be supplemented with the basic object detection result. Thirdly, the quality of each candidate bounding box will be evaluated, and a K-means based vertex clustering algorithm will be used to optimize the qualified candidate bounding boxes. Finally, based on the semantic segmentation results, the contours of the scene elements belonging to the stuff class will be retained, and the contours of the scene elements belonging to the things class will be matched with candidate bounding boxes.

Since the perception results obtained by the proposed algorithm are similar to panoptic segmentation algorithms, PQ [2] is also suitable to evaluate the performance of the proposed algorithm. On the Cityscapes dataset [29], three groups of experiments are carried out: upper limit verification experiments, lower limit verification experiments, and cross verification experiments. The experimental results show that the precise scene perception results can be obtained by the proposed method without additional complex instance-level data annotation.

In addition, the basic semantic segmentation results contribute more to the final perception effect than the basic object detection results.

The main contributions of this paper are summarized as follows:

(1) A joint perception method is proposed, which can obtain the perception results similar to panoptic segmentation through combining object detection and semantic segmentation, and can achieve both practicability and high accuracy;
(2) A grid based contour vertex clustering algorithm is designed to iteratively refine the candidate bounding box, which can extract the backbone of each scene element from noises;
(3) The feasibility of the proposed method is verified by the upper limit verification and the lower limit verification experiments, and the main impact of the final perception effect is clarified by the cross validation experiments which are the basic semantic segmentation results.

The rest of this paper is organized as follows. In Sect. 2, some preliminaries and related work will be introduced. Section 3 presents the principle and implementation details of the proposed method. Experimental results and analyses to demonstrate the effectiveness of the method are arranged in Sect. 4. Section 5 discusses some concluding remarks of this paper and the prospects of future work.

2 Related Work

Scene perception is also known as scene parsing [1], image parsing [30], or holistic scene understanding [31]. For the visual perception of traffic scenes, various algorithms have been proposed for identifying pedestrians, vehicles and roads. Before the widespread use of deep learning, traditional methods focused on specific target information perception in traffic scenes. Compared with traditional methods, deep learning algorithms for visual perception have more advantages. The deep neural networks for single-tasks can greatly improve the perception performance. And multi-task deep neural networks can also achieve amazing scene perception performance and understand complete traffic scene information at the same time.

2.1 Traditional Scene Perception Algorithms

Traditional algorithms for traffic scene perception can be categorized into two classes: image processing based algorithms and machine learning based algorithms. Image processing based perception algorithms normally use basic image operations to realize scene perception. These basic image operations include image pre-processing (e.g., smoothing, sharpening and histogram equalization), image space conversion (e.g. color space conversion and space domain transformation), edge detection (e.g., canny operator and sobel operator), morphological operations (e.g., dilation, erosion and skeleton) and image segmentation (e.g., flood-fill and graph-cut). By synthesizing various image processing operations, Fan et al. [32] proposed a real-time lane detection algorithm using binocular stereo vision. Machine learning based perception algorithms normally adopt a combination framework with a front-end image feature extractor and a back-end object classifier. Traditional feature extractors include Haar-like, HOG, LBP and SIFT. Decision tree, Bayesian network, artificial neural network, and support vector machine can be used as object classifiers. To achieve good results, the analysis of the specific problems is usually carried out as the first procedure in the traditional methods.

However, the robustness of the traditional methods is difficult to be guaranteed. Yao et al. [31] combined traditional object detection, scene classification and semantic segmentation to transform the whole problem into predictions of primitive structure to achieve holistic scene perception. However, traditional methods based traffic scene perception systems are large and cumbersome, and the final performance and the robustness cannot be achieved as expected.

2.2 Scene Perception Deep Neural Networks

Entity-level object detection is the most common technology for scene perception tasks. With the development of deep learning algorithms, a series of significant breakthroughs have been made in the field of multi-object detection. Faster R-CNN proposed by Ren et al. [3] could perceive a lot of elements in traffic scenes. However, a large number of candidate objects need to be verified in this kind of two-stage neural networks for object detection, which makes the overall efficiency unsatisfying. Therefore, single-stage object detection neural networks such as RetinaNet [4], SSD [5], YOLO [6] and CenterNet [7] were proposed for the efficiency issue. Although rectangular detection boxes are suitable for most scene elements, they can hardly describe the specific contours of the elements. Moreover, only detecting scene elements without identifying their attributes leaves much to be desired. It's also a big challenge to achieve high accuracy of traffic scene perception due to the unavoidable false positive and false negative results.

Pixel-level semantic segmentation plays a very important role in scene visual perception. Thanks to the development of deep learning, researchers have made great breakthroughs in semantic segmentation. a variety of derivative and variant neural networks have emerged since FCN [8] was proposed as the first deep neural network for image semantic segmentation. Papers [9] and [15] made reviews on many semantic segmentation oriented deep neural networks. UNET [10], Tiramisu [11], SegNet [12], DeepLab [13] and BiSeNet [14] are the most representative architectures of neural networks for semantic segmentation. Yang et al. [33] discussed some defects of object detection based scene understanding methods, and proposed a semantic segmentation based neural network model for terrain perception which can be used as an assistant for blindman navigation. Zhou et al. [34] proposed a real-time perception method for drivable path prediction based on semantic segmentation. Although the contours of scene elements can be identified clearly, the labels of pixels are not always correct. In addition, when scene elements belonging to the same kind gather together to form a group, the individuals in the group cannot be identified successfully.

In order to overcome the inherent defects of individual perception technology, multi-task perception is becoming a trend. Teichmann et al. [35] proposed a multi-task learning network named MultiNet, which combined object detection and semantic segmentation. In the form of a unified neural network, MultiNet dealt with the task of object detection, object classification and semantic segmentation, and achieved scene perception in real time. It was found that the performance of multi-task learning neural network is closely related to the weight distribution where the loss value of each branch task is calculated. Therefore, Kendall et al. [36] proposed a multi-task learning method using uncertainty to weigh losses, which achieved advanced performance in scene understanding. Hu et al. [37] proposed a multi-task neural network based on Faster R-CNN to complete vehicle identification and body attribute prediction simultaneously. Similar to Mask R-CNN, Dvornik et al. [38] designed a deep neural network structure, BlitzNet, for real-time scene understanding. Since most of the current algorithms are trained individually for each procedure in traffic scene recognition

tasks, Cheng et al. [39] proposed an end-to-end multi-task neural network, Dense-ACSSD, to implement multi-object detection and drivable area segmentation.

Sometimes, multi-task perception results can be improved by fusion processing. Panoptic segmentation can achieve the most advanced scene perception performance by fusing semantic segmentation results and instance segmentation results. The main variations of the existing neural networks for panoptic segmentation lie in two aspects: the acquisition of basic semantic segmentation and instance segmentation results, and the fusion algorithm of the two results. Among them, the acquisitions of basic results are quite similar, while the fusion of basic results shows some differences. Panoptic segmentation deep neural networks, such as PFPNet [17], AdaptIS [18], Seamless [20], TASCNet [22], AUNet [23], DeeperLab [24], Panoptic FCN [26], MaX-DeepLab [27], EfficientPS [28], simply fused the results of semantic segmentation and instance segmentation under the artificial heuristic rules. UPSNet [21] introduced a parameter-free panoptic decoder to fuse the basic results through pixel classification. OANet [25] designed a spatial sorting module to achieve fusion. Inspired by the representation of scene graphs, SOGNet [19] used categories, geometries and attributes of each scene element to perform a unified spatial embedding representation, and guided specific fusion processing by modeling overlap relations among instances. Panoptic FCN [26] present a conceptually simple, strong, and efficient framework for panoptic segmentation which aims to represent and predict foreground things and background stuff in a unified fully convolutional pipeline. MaX-DeepLab [27] designed an end-to-end model for panoptic segmentation which simplifies the current pipeline that depends heavily on surrogate sub-tasks and hand-designed components, such as box detection, non-maximum suppression, thing-stuff merging, etc. EfficientPS [28] architecture that consists of a shared backbone which efficiently encodes and fuses semantically rich multi-scale features.

Different from panoptic segmentation, a joint perception method is proposed in this paper which combines the results of basic object detection and semantic segmentation. Similar to TASCNet known as a panoptic segmentation algorithm, the proposed method also uses binary mask image processing but with different specific processing targets. The proposed method is different from the method in paper [1] by combining the results of basic object detection and semantic segmentation. To be specific, the proposed method focuses on the parallel fusion of basic perceptual results based on deep learning, while the method in paper [1] focuses on cascade semantic inference based on traditional methods. Since the accuracy of the object detection results is lower than that of the instance segmentation results, candidate bounding boxes are optimized to be matched with more detailed contours in the proposed method.

3 The Method

In this paper, a joint object detection and semantic segmentation based traffic scene perception method is proposed. The overview of the method is shown in Fig. 1. Firstly, the basic object detection and semantic segmentation results are obtained by deep neural networks. Secondly, under the principle of information entropy, basic perception results are preprocessed. Thirdly, candidate bounding boxes are supplemented by the joint algorithm. Fourthly, the quality of each candidate bounding box is evaluated and further optimized within a manageable range. Finally, contour matching is applied to achieve perception results which can be similar to that achieved by panoptic segmentation methods.

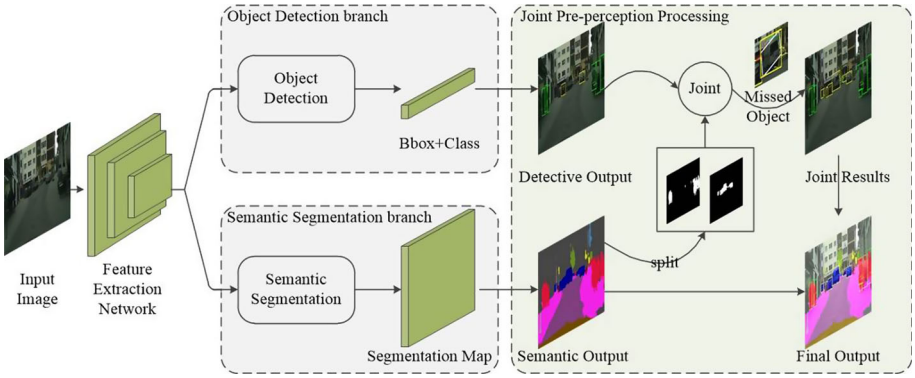

Fig. 1 Overview of the proposed joint pre-perception algorithm for mobile traffic scenes

3.1 Basic Perceptual Information Capturing

Basic perceptual information of traffic scenes includes object detection results and semantic segmentation results. There are two key ways to obtain perceptual information, one is using two independent deep neural network models for each task, the other is training one multi-task neural network model for both tasks. The same training data should be prepared for these two ways. The scene object detection task needs entity-level annotations in the form of a bounding box, while the scene semantic segmentation task requires pixel-level category annotations. Obviously, compared with the data annotation required by panoptic segmentation, the data annotations required by simple object detection and semantic segmentation are simpler and more practical. The concrete perception models are introduced as follows.

The output of the object detection model is a series of bounding boxes and the corresponding category numbers of detected objects. For an input image X, the goal is to maximize the likelihood probability $p_{\mathrm{d}}(O|X)$ of the target sequence $O = \{o_i | i = 1, 2, \ldots, n_{\mathrm{o}}\}$ and the confidence level of each object o_i, where o_i is usually composed of a tuple (c, x, y, w, h, t), and c represents the category of the object, x、y、w and h represent the abscissa and ordinate of the bounding box center, and the width and height of the bounding box respectively. Note x、y、w and h are actually relative values, t represents the confidence level ($t \in [0,1]$) when o_i is predicted as category c, and n_{o} is the number of detected objects. The training task can be described as maximizing the probability of the target sequence by solving Eq. (1):

$$O^* = \arg\max_{\theta} \sum_{X,O} \log p_{\mathrm{d}}(O|X;\theta) \tag{1}$$

where θ represents the weight of the neural network model, and O^* is a predicted sequence derived from the input image X while the optimal training weight θ^* is used. It should be noted that the form of tuple o_i in this paper is only one of the major forms.

The output of the semantic segmentation model is a scene semantic segmentation image. For an input image X, the goal is to assign a unique label to each pixel (i, j) with the category set $c_{i,j}(c_{i,j} = 1, \ldots, n_{\mathrm{c}})$, and to output a scene semantic segmentation image Y, where n_{c} is the number of categories of the semantic labels. The overall training process of semantic segmentation model is shown in Eq. (2):

$$Y^* = \arg\max p_{\mathrm{s}}(y_{i,j} = c_{i,j}|X;w) \tag{2}$$

where w represents the weight parameter of semantic segmentation neural network, p_s is the predicted distribution probability, $y_{i,j}$ is the predicted value of the pixel (i, j) in the output image Y corresponding to the weight w, and Y^* is the predicted result derived from the optimal weight w^*.

3.2 Perceptual Information Pre-Processing

Perceptual information preprocessing aims to analyze the topological structure of the semantic output image Y^* to obtain locations of scene objects, or object groups, which belong to the things class.

In this method, the image boundary tracking algorithm is used for semantic analysis. Structured information can be obtained by the boundary tracking algorithm encoding the boundary contour of each object into a chain code with a series of points. Since the scene semantic segmentation image Y^* is too complex to be analyzed, Y^* will be divided into a binary image sequence $\{Y_i | i = 1, \ldots, n_c\}$ according to the pixel value. As shown in Fig. 2, the original semantic segmentation image is split into a series of binary images, and each image represents one category. Assuming that event A in scene perception means that the contour of one scene object is clearly defined, and $p(A)$ represents the probability of random event A. Then the information entropy $H(A)$ of the scene semantic perception task can be formulated as Eq. (3), where A_i represents the event that the i-th object in the scene can match its contour, and n is the number of objects. Under the principle of information entropy, the smaller the probability of an event occurring, the more information can be achieved. To obtain more information, attention should be focused on those objects whose contours are difficult to be obtained.

$$H(A) = -\sum_{i=1}^{n} p(A_i) \log(p(A_i)) \tag{3}$$

Based on information entropy theory, the scene elements belonging to the things class (pedestrians and vehicles) with dense semantic contour have higher priority to be obtained.

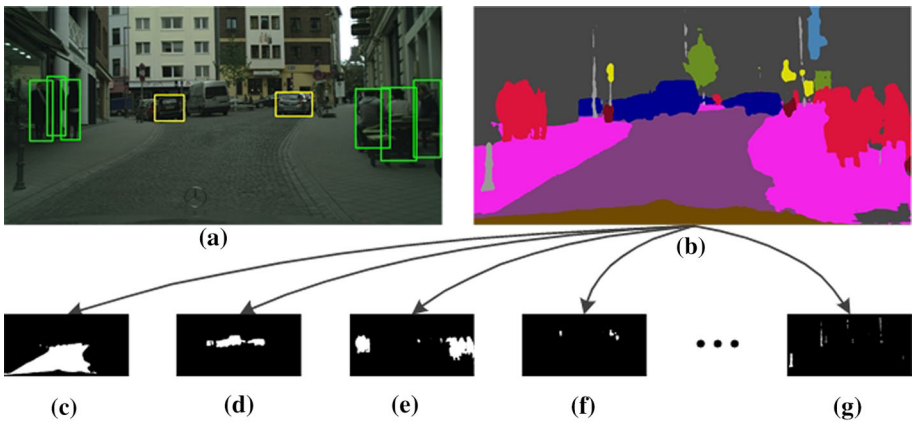

Fig. 2 Examples of Semantic image splitting: **a** raw image with detected information; **b** semantic segmentation image; **c** binary image of the road; **d** binary image of vehicles; **e** binary image of pedestrians; **f** binary image of traffic signs; **g** binary image of poles

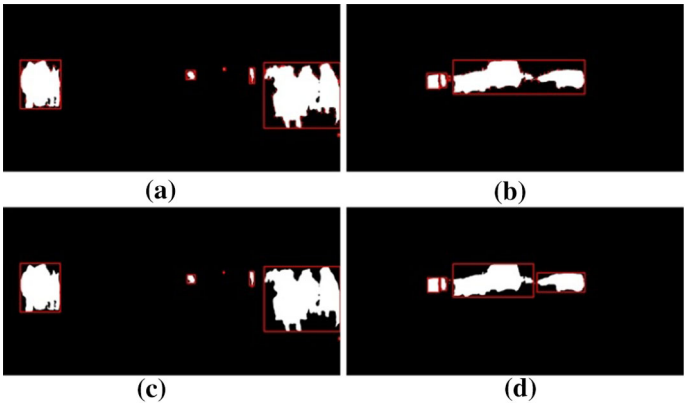

Fig. 3 Examples of semantic image processing: **a** semantic pedestrian detection with noise; **b** semantic vehicle detection with noise; **c** semantic pedestrian detection resulted from the algorithm; **d** semantic vehicle detection resulted from the algorithm

Specifically, according to the pixel category, Y^* is scanned to form vehicle mask image $Y_v = \{y_{i,j} \in Tag_v\}$ and pedestrian mask image $Y_p = \{y_{i,j} \in Tag_p\}$, where $y_{i,j}$ represents the category of the pixel (i, j) in Y^*, Tag_v and Tag_p are the sets of vehicle related tags (car, bus, truck, etc.) and pedestrian related tags (man, woman, child, etc.) respectively. Then image scaling is applied to improve the computational efficiency, and morphology close operation is used to denoise the mask images. Therefore, Y_v is enhanced to $Y_v{}'$, and Y_p is enhanced to $Y_p{}'$. Afterwards, a boundary tracking algorithm is applied to $Y_v{}'$ and $Y_p{}'$, and two output sets are obtained, i.e., the set of vehicle contours $D_v = \{d_v{}^i | i = 1,2,\ldots,n_v\}$, and the set of pedestrian contours $D_p = \{d_p{}^i | i = 1,2,\ldots,n_p\}$, where $d_v{}^i$ represents the contour of the i-th vehicle, $d_p{}^i$ represents the contour of the i-th pedestrian, and n_v and n_p are the total number of vehicles and pedestrians with contours respectively. Finally, the minimum rectangular bounding box of each object in D_v and D_p is calculated according to its contour, and two new sets are formulated as $D_v{}' = \{v_d{}^i = box(d_v{}^i) | i = 1, 2, \ldots, n_v\}$, $D_p{}' = \{p_d{}^i = box(d_p{}^i) | i = 1, 2, \ldots, n_p\}$, where $box(*)$ is a function for the bounding rectangle calculation. Figure 3 shows the actual effect of mask image preprocessing. Locations of candidate vehicles and pedestrians have been perceived successfully.

3.3 Perceptual Information Combination

Perception information combination is based on the semantic perception results $D_v{}'$ and $D_p{}'$, and the object detection results O^*. Before the combination procedure, O^* is divided into a set of candidate vehicle $O_v = \{v_o{}^j = o_i | o_i \in Tag_v, j = 1, 2, \ldots, n_v{}'\}$ and a set of candidate pedestrians $O_p = \{p_o{}^j = o_i | o_i \in Tag_p, j = 1, 2, \ldots, n_p{}'\}$, where $n_v{}'$ and $n_p{}'$ are the numbers of detected vehicles and pedestrians respectively. More details about the enhancement of the perception effect by combination are as follows.

The pairing matching algorithms can be used to filter the false positives. In order to filter false positives, elements from both $D_v{}'$ and O_v are unified into the same image coordinate system. Then, every object in $D_v{}'$ is compared to the objects in O_v. For $v_d{}^i \in D_v{}'$ and $v_o{}^j \in O_v$, if $IOU(v_d{}^i, v_o{}^j) > T_{iou}$, the pair of $(v_d{}^i, v_o{}^j)$ will be added to the candidate vehicle queue V, where $IOU(*)$ is the function to calculate the special relationship of two rectangular boxes

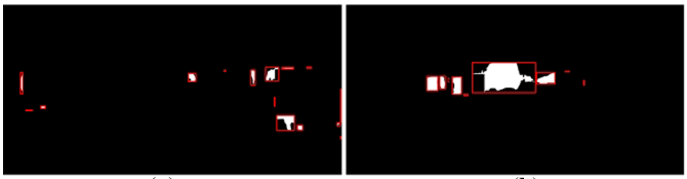

Fig. 4 Left candidate regions of semantic object: **a** candidate regions of pedestrians; **b** candidate regions of vehicles

and T_{iou} is a threshold parameter. Since there is a one-to-one correspondence between $v_{\text{d}}{}^{i}$ from $D_{\text{v}}{}'$ and $d_{\text{v}}{}^{i}$ from D_{v}, for each pair of elements $(v_{\text{d}}{}^{i}, v_{\text{o}}{}^{j})$ in V, a new set $G_{\text{v}} = \{g_{\text{v}}{}^{j} | j = 1, 2, \ldots, n_{\text{gv}}\}$ can be constructed with $g_{\text{v}}{}^{j} = \{(x_{\text{p}},y_{\text{p}}) | (x_{\text{p}},y_{\text{p}}) \in d_{\text{v}}{}^{i}$ and $\varphi((x_{\text{p}},y_{\text{p}}), v_{\text{o}}{}^{j}) = 1\}$ consisted of the points in both region $v_{\text{o}}{}^{j}$ and contour $d_{\text{v}}{}^{i}$, here $(x_{\text{p}},y_{\text{p}})$ represents the coordinate value of the boundary point, $\varphi((x_{\text{p}},y_{\text{p}}), v_{\text{o}}{}^{j})$ is an indicative function of whether the calculated point $(x_{\text{p}},y_{\text{p}})$ is in region $v_{\text{o}}{}^{j}$, and n_{gv} is the number of elements of the set G_{v}. Meanwhile, $v_{\text{o}}{}^{j}$ is added to the final bounding box set of vehicles V_{f}. It is worth noting that the pairing matching operation not only filters the false positives in object detection results, but also gets the preparation work of recalling the false negatives ready.

Based on the pairing matching results, the false negatives can be recalled by finding missing objects from the pre-processed mask images. First of all, for the mask image $Y_{\text{v}}{}'$, mask regions are removed if it appears in V_{f} by setting the pixel value to 0. Secondly, for the new image $Y_{\text{v}}{}'$, a boundary tracking algorithm is applied again to obtain the supplementary candidate vehicle contour set $D_{\text{w}} = \{d_{\text{w}}{}^{i} | i = 1, 2, \ldots, n_{\text{vw}}\}$, where $d_{\text{w}}{}^{i}$ represents the contour of the i-th candidate vehicle, and n_{vw} is the total number of supplementary candidate vehicles. Thirdly, rectangular bounding boxes of contours in D_{w} are calculated, and constitute the set $V_{\text{w}} = \{v_{\text{w}}{}^{i} | i = 1, 2, \ldots, n_{\text{vw}}\}$. As shown in Fig. 4, the recalled false negatives might introduce new noises, which will reduce the accuracy of the recall results.

In order to filter the noises and extract the trunks, the recalled false negatives need to be post-processed. Specifically, the recalled results with a rather small area are filtered at first. Then, based on the set V_{w}, for every $v_{\text{w}}{}^{i} \in V_{\text{w}}$, the proportion of the pixels that belong to the vehicle category in the total rectangular area is calculated as $r_{\text{w}}{}^{i}$. If $r_{\text{w}}{}^{i} > T_{\text{rate}}$, $v_{\text{w}}{}^{i}$ is added to set V_{f}, the final bounding box set of vehicles; otherwise, the current bounding box is not ideal since a large number of irrelevant pixels are included.

For the latter case mentioned above, a grid based contour vertex clustering algorithm is designed to iteratively refine the candidate bounding box. In the image coordinate system, the current region $v_{\text{w}}{}^{i}$ is divided into four sub-regions by a grid. Thus, the number of clusters to be clustered by the K-means algorithm is set to "4", and each cluster center is initialized as the center of the corresponding grid cell. Then, the Euclidean distance between boundary points is adopted as the distance measurement of two clusters, and the final center points of four clusters are named as $p_{\text{c1}}, p_{\text{c2}}, p_{\text{c3}}$ and p_{c4}. Afterwards, the external bounding box noted as $v_{\text{w}}{}^{i*}$ is calculated and then added to the set V_{f}. Since there is a one-to-one correspondence between $v_{\text{w}}{}^{i*}$ and $d_{\text{w}}{}^{i}$ from D_{w}, the set G_{v} is supplemented with $g_{\text{v}} = \{(x_{\text{p}},y_{\text{p}}) | (x_{\text{p}},y_{\text{p}}) \in d_{\text{w}}{}^{i}$ and $\varphi((x_{\text{p}},y_{\text{p}}), v_{\text{w}}{}^{i*}) = 1\}$. As shown in Fig. 5, although the original recalled false negative result is rough, the refined result from the proposed K-means based vertex clustering algorithm is much more detailed.

After the above joint algorithm, set V_{f} stores the joint enhanced results of vehicular bounding boxes, set G_{v} records the results of vehicular contour, and the elements of both V_{f}

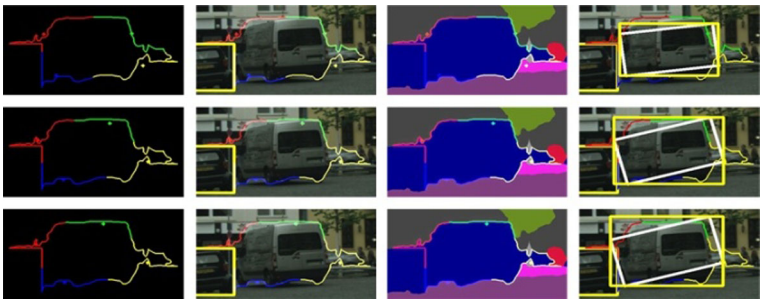

Fig. 5 K-Means based semantic principal component region extraction (the first row: iteration = 0; the second row: iteration = 3; the third row: iteration = 10)

and G_v are one-to-one mapped. By using the same algorithm, the joint enhanced results of the pedestrian can also be obtained and stored in P_f and G_p. In Fig. 6, each line is a scene sample showing the comparison between the results obtained before processing and achieved from the joint pre-perception algorithm respectively. The use of the joint pre-perception algorithm can remove the false negative objects and recall the false negatives ones.

Finally, the perception results of the panoptic segmentation style can be obtained after scene element contour matching. Specifically, according to the one-to-one relationship between V_f and G_v, with the correspondence between P_f and G_p, the instance-like segmentation perception results can be generated. Obviously, the utilize of heuristic rules by fussing the perceptual results of instance-like segmentation and basic semantic perceptual information can provide similar results as panoptic segmentation. As shown in Fig. 7, scene

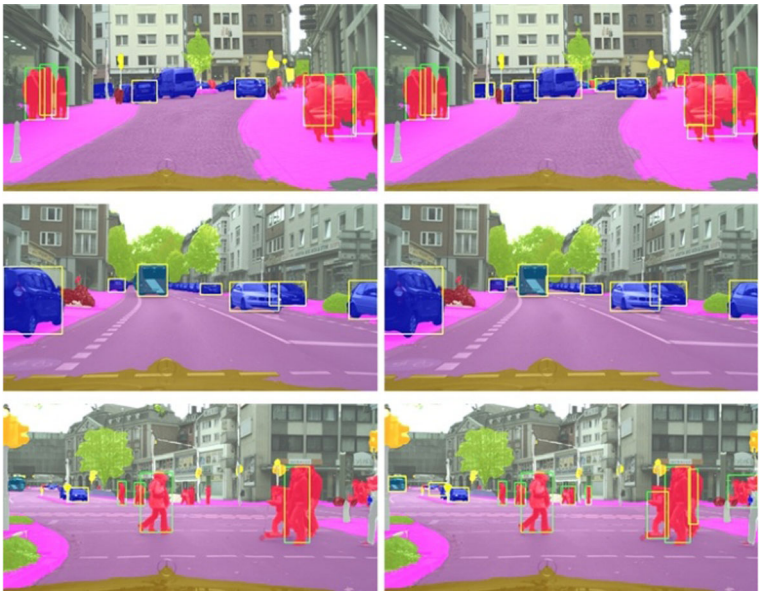

Fig. 6 The necessity of joint pre-perception (the first column and the second column represent results obtained before processing and results achieved from the joint pre-perception algorithm respectively)

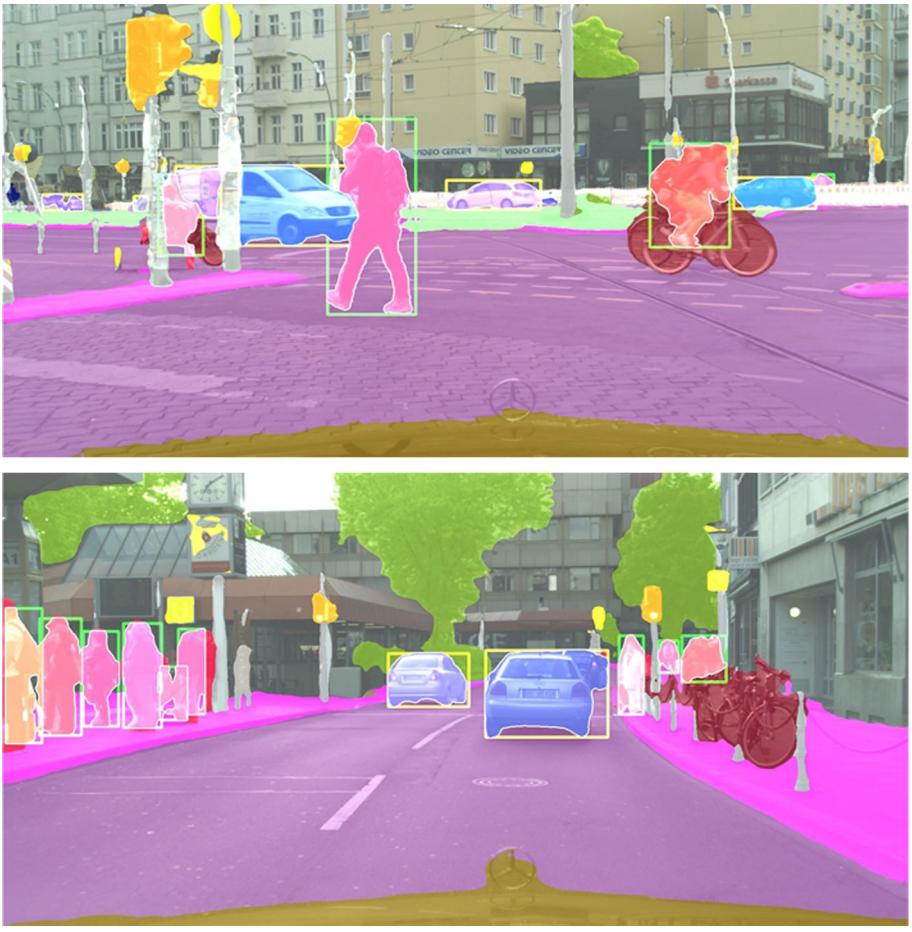

Fig. 7 Panoptic-like segmentation examples

elements of both the stuff class and the things class are presented in a panoptic segmentation style.

4 Experiments

The proposed joint perception algorithm which effectively combines the results of object detection and semantic segmentation can achieve a similar effect to panoptic segmentation. Therefore, three criteria defined for panoramic segmentation in paper [2] are used for evaluation, i.e., *PQ* (panoptic segmentation), *SQ* (segmentation quality) and *RQ* (recognition quality). The formulations are shown in Eq. (4):

$$\begin{cases} SQ = \dfrac{\sum_{(p,g\in TP)} IOU(p,g)}{|TP|} \\ RQ = \dfrac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \\ PQ = \dfrac{\sum_{(p,g\in TP)} IOU(p,g)}{|TP|} \times \dfrac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \end{cases} \tag{4}$$

where p is the predicted result, g is the ground-truth, TP is the positive result, FP is the false positive result, FN is the false negative result, $IOU(p, g)$ is the function for evaluating the proportion of the pixels that intersection of p and g over union of p and g. RQ is widely used in object detection known as F1-score, which is used to calculate the accuracy of object recognition for each element in perception. SQ is the average intersection ratio of predicted semantic segmentation results and ground-truth. As discussed in paper [2], $PQ = SQ \times RQ$ can provide insight for analysis which measures performance of all classes in a uniform way using a simple and interpretable formula. Generally, $IOU(p, g) > 0.5$ is regarded as the matching condition.

In order to prove the effectiveness and accuracy of the proposed method, three groups of experiments are implemented: upper limit verification, lower limit verification, and cross verification. The upper limit verification is carried out when the results of semantic segmentation and object detection are completely consistent with the actual annotations, i.e., the ground-truth is used to verify the proposed joint perception algorithm. The lower limit verification is carried out when the two basic perception results are poor, i.e., perception models are trained on different datasets. The additional cross validation is designed to evaluate how basic perceptual information makes an influence on the final performance.

In order to facilitate the comparison with the panoramic segmentation methods, the proposed method is tested on the verification set of the Cityscapes dataset. All test methods reach the following agreement conditions: 19 kinds of targets in the original data annotation are used, including 11 kinds of stuff objects (road, sidewalk, building, wall, fence, pole, trafficlight, trafficsign, vegetation, terrain and sky) and 8 kinds of things objects (person, rider, car, truck, bus, train, motorcycle, cycle). Specifically, the upper limit verification uses the pixel level semantic segmentation annotation and the bounding box level target detection annotation directly, and the lower limit verification uses two models opened by OpenVINO: semantic segmentation adas-0001 and person vehicle bike detection cross road-0078. The model of object detection for them is obtained from the data training of the fixed-point traffic scene, which meets the requirements of the perception model in the lower limit evaluation for the traffic scene. Cross validation uses the two remaining combinations, A + O and O + A, shown in Table 1.

As shown in Table 1, the results of the upper limit validation experiment illustrate that the theoretical upper limit approximates the ground-truth of panoptic segmentation. In other words, when the basic semantic segmentation and object detection results are perfect, the proposed joint algorithm works well. When the basic semantic segmentation and object detection results are quite poor, the lower limit is still acceptable even though there is a gap between the results of the proposed algorithm and that of the state-of-the-art. As shown in Table 2, in the upper limit validation experiment, the PQ values are quite high in general, but not ideal in some semantic categories. Specifically, the PQ value is relatively low due to the high probability of occlusion in the categories like pedestrian and vehicle of which the number is large. As shown in Table 3, in the lower limit verification experiment, the SQ

Table 1 Comparison of experimental results on Cityscapes val Benchmark

Method	Backbone	PQ	SQ	RQ	Miou	PC
Panoptic (Merge) [17]	–	61.2	80.9	74.4	–	–
AdaptIS [18]	ResNet-101	60.6	–	–	79.3	–
SOGNet [19]	ResNet-50	60.0	–	–	–	–
Seamless [20]	ResNet-50	59.8	–	–	75.4	–
UPSNet [21]	ResNet-50	59.3	79.7	73.0	75.2	–
TASCNet [22]	ResNet-101	59.2	–	–	77.8	–
AUNet [23]	ResNet-101	59.0	–	–	75.6	–
Panoptic FPN [13]	ResNet-101	58.1	–	–	75.7	–
DeeperLab [24]	Xception-71	56.5	–	–	–	75.6
Ours (upper limit)	Annotation	95.4	97.5	97.6	95.7	–
Ours (lower limit)	OpenVINO	37.1	73.8	47.4	36.7	–
Ours (mixed)	A + O	79.4	94.0	83.0	76.2	–
Ours (mixed)	O + A	39.5	74.6	50.4	39.0	–

Table 2 Experimental results on Cityscapes val Benchmark (upper evaluation)

Element	Type	PQ	SQ	RQ	mIOU
Road	Stuff	100	100	100	100
Sidewalk	Stuff	100	100	100	100
Building	Stuff	100	100	100	100
Wall	Stuff	100	100	100	100
Fense	Stuff	99.5	100	99.5	98.9
Pole	Stuff	100	100	100	100
Trafficlight	Stuff	100	100	100	100
Trafficsign	Stuff	100	100	100	100
Vegetation	Stuff	100	100	100	100
Terrain	Stuff	100	100	100	100
Sky	Stuff	100	100	100	100
Person	Thing	77.1	89.7	85.9	74.9
Rider	Thing	92.9	95.9	96.9	93.9
Car	Thing	70.8	84.0	84.3	72.9
Truck	Thing	97.5	98.6	100	97.9
Bus	Thing	98.1	98.1	100	100
Train	Thing	99.4	99.4	100	100
Motorcycle	Thing	90.2	94.0	96.0	92.3
Bicycle	Thing	86.3	93.0	92.9	86.7

PS: mIOU represents the average IOU of a single category

Table 3 Experimental results on Cityscapes val Benchmark (lower evaluation)

Element	Type	PQ	SQ	RQ	mIOU
Road	Stuff	92.6	92.6	99.7	99.4
Sidewalk	Stuff	59.9	77.6	77.2	62.8
Building	Stuff	82.6	85.5	96.6	93.5
Wall	Stuff	20.9	72.9	28.7	16.8
Fense	Stuff	17.9	69.7	25.7	14.8
Pole	Stuff	19.1	59.4	32.1	19.1
Trafficlight	Stuff	28.2	62.7	44.9	28.9
Trafficsign	Stuff	40.0	69.2	57.8	40.7
Vegetation	Stuff	83.3	86.1	96.7	93.7
Terrain	Stuff	13.4	63.5	21.1	11.8
Sky	Stuff	79.0	89.8	88.0	78.5
Person	Thing	20.1	68.1	29.6	17.3
Rider	Thing	21.6	62.7	34.5	20.9
Car	Thing	32.9	76.7	42.9	27.3
Truck	Thing	17.9	83.7	21.4	12.0
Bus	Thing	31.4	81.7	38.4	23.8
Train	Thing	15.6	70.0	22.2	12.5
Motorcycle	Thing	11.1	65.7	16.9	9.2
Bicycle	Thing	16.6	65.0	25.6	14.7

PS: mIOU represents the average IOU of a single category

values of the proposed method are not low, but poor *RQ* values could lead to unsatisfactory *PQ* values. This means that there are a lot of false positives and false negatives in the perception results which should blame on the basic semantic segmentation results. In addition to the indicators, Fig. 7 shows the panoptic-level segmentation results in the lower limit verification experiment. Obviously, the scale of the scene element is an important factor that affects the perceived performance. As shown in Table 1, through additional mixed experiments, it can be found that when the result of semantic segmentation is perfect and even the result of object detection is not good, the proposed algorithm can still maintain a high performance; while when the result of semantic segmentation is not good and the result of object detection is perfect, the actual results of the proposed joint algorithm is close to that of the lower limit evaluation. Apparently, the result of the basic semantic segmentation determines the theoretical lower limit of perception performance, while the result of basic object detection determines the theoretical upper limit of perception performance. In general, the proposed joint perception algorithm is effective and has high accuracy.

5 Conclusion

In this paper, a joint perception algorithm for traffic scenes is proposed, which combines object detection and semantic segmentation. This is an attempt at holistic traffic scene parsing. Under the principle of information entropy, the perception of pedestrians and vehicles

is targeted. Through the flexible application of image processing technology, the joint perception algorithm achieves panoptic-level segmentation performance. The proposed method can take both the accuracy and practicality into account without complex data annotation as panoptic segmentation required. Competitive performance is achieved on the Cityscapes dataset, and the importance of basic semantic segmentation results during the joint progress is verified. Since there is no need to specify the basic perception model, the algorithm has a wide range of generality. However, compared to the state-of-the-art panoptic segmentation technology, the proposed method still has defects in the effect of instance segmentation. Since the proposed method is based on both results of the basic semantic segmentation and object detection, either mistakes will lead to undesirable results. And occlusions or similar traffic scene elements also may lead to errors during semantic boundary calculation of adjacent elements. Thus, there are still improvements to be made in traffic scene perception algorithms based on joint object detection and semantic segmentation.

Acknowledgements This work is being supported by the National Key Research and Development Project of China under Grant No. 2020AAA0104001, the Zhejiang Lab. under Grant No. 2019KD0AD011005 and the Zhejiang Provincial Natural Science Foundation of China under Grant No. LQ22F020008.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Tighe J, Niethammer M, Lazebnik S. (2014) Scene parsing with object instances and occlusion ordering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3748–3755
2. Kirillov A, He K, Girshick R, et al. (2019) Panoptic segmentation. 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), long beach, CA, USA, 2019, pp 9396–9405
3. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Machine Intell. https://doi.org/10.1109/TPAMI.2016.2577031
4. Lin T, Goyal P, Girshick R, et al. (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision pp 2980–2988
5. Liu W, Anguelov D, Erhan D et al (2016) SSD: single shot multibox detector. In: Leibe Bastian, Matas Jiri, Sebe Nicu, Welling Max (eds) Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I. Springer International Publishing, Cham, pp 21–37
6. Redmon J, Divvala S, Girshick R, Farhadi A. (2016) You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
7. Duan K, Bai S, et al. (2019) Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6569-6578
8. Long J, Shelhamer E, Darrell T. (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440
9. Garcia-garcia A, Orts-escolano S, Oprea S et al (2018) A survey on deep learning techniques for image and video semantic segmentation. Appl Soft Comput 70:41–65
10. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. international conference on medical image computing and computer assisted intervention, Munich, Germany, pp 234–241
11. Jégou S, Drozdzal M, Vazquez D, Romero A, Bengio Y. (2017) The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 11–19
12. Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: A deep convolu- tional encoder-decoder architecture for image segmentation. IEEETrans Pattern Anal Mach Intell 39(12):2481–2495

13. Chen L, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018) DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Machine Intel 40(4):834–848
14. Yu C, Wang J, Peng C, et al. (2018) BiSeNet: bilateral segmentation network for real-time semantic segmentation. European Conference on computer vision, Munich, Germany, 2018, pp.334–349.
15. Siam M, Gamal M, Abdel-Razek M, et al. (2018) A comparative study of real-time semantic segmentation for autonomous driving. In: IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), Salt Lake City, UT, pp 700–710.
16. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. (2017) In: Proceedings of the IEEE international conference on computer vision, Venice, Italy, pp 2961–2969
17. Kirillov A, Girshick R, He K, Dollár P. (2019) Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Long Beach, CA, USA, pp 6399–6408
18. Sofiiuk K, Barinova O, Konushin A. (2019) Adaptis: Adaptive instance selection network. In: Proceedings of the IEEE/CVF international conference on computer vision, Seoul, Korea pp 7355–7363
19. Yang Y, Li H, Zhao Q et al (2020) SOGNet: scene overlap graph network for panoptic segmentation. Proc AAAI Conf Artif Intel 34(07):12637–12644
20. Porzi L, Bulò SR, Colovic A, Kontschieder P (2019) Seamless scene segmentation. 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), Long Beach, CA, USA, pp 8269–8278
21. Xiong Y, Liao R, Zhao H, Hu R, Bai M, Yumer E, Urtasun R. (2019) Upsnet: a unified panoptic segmentation network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8818–8826
22. Li J, Bhargava A, Tagawa T et al.(2019) Learning to Fuse Things and Stuff. arXiv: 1812.01192v2,
23. Li Y et al. (2019) Attention-guided unified network for panoptic segmentation. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), Long Beach, CA, USA, 2019, pp 7019–7028
24. Yang TJ, Collins MD, Zhu Y, Hwang JJ, Liu T, Zhang X, Sze V, Papandreou G, Chen LC. (2019) Deeperlab: Single-shot image parser. arXiv preprint arXiv:1902.05093
25. Liu H, Peng C, Yu C, Wang J, Liu X, Yu G, Jiang W. (2019) An end-to-end network for panoptic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp 6172–6181
26. Li Y, Zhao H, Qi X, et al. (2021) Fully convolutional networks for panoptic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 214–223
27. Wang H, Zhu Y, Adam H, Yuille A, Chen LC. (2021) Max-deeplab: End-to-end panoptic segmentation with mask transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp 5463–5474
28. Mohan R, Valada A (2021) Efficientps: Efficient panoptic segmentation. Int J Comput Vision 129(5):1551–1579
29. Cordts M, Omran M, Ramos S, et al. (2016) The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition pp 3213–3223
30. Tu Z, Chen X, Yuille AL, Zhu SC et al. (2003) Image parsing: unifying segmentation, detection, and recognition. In: Proceedings ninth IEEE international conference on computer vision, Nice, France, pp 18–25
31. Yao J, Fidler S, Urtasun R. (2020) Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In: 2012 IEEE conference on computer vision and pattern recognition, Providence, RI, pp 702–709. IEEE.
32. Fan R, Dahnoun N (2018) Real-time stereo vision-based lane detection system. Measure Sci Technol 29(7):074005
33. Yang K, Wang K, Zhao X, Cheng R, Bai J, Yang Y, Liu D (2017) IR stereo RealSense: decreasing minimum range of navigational assistance for visually impaired individuals. J Ambient Intel Smart Environ 9(6):743–755. https://doi.org/10.3233/AIS-170459
34. Zhou W, Worrall S, Zyner A, Nebot E. (2018) Automated process for incorporating drivable path into real-time semantic segmentation. In: 2018 IEEE International Conference on Robotics and Automation (ICRA) Brisbane, QLD pp 6039–6044. IEEE
35. Teichmann M, Weber M, Zoellner M, Cipolla R, Urtasun R. (2018) Multinet: Real-time joint semantic reasoning for autonomous driving. In: 2018 IEEE intelligent vehicles symposium (IV) Changshu pp. 1013–1020
36. Kendall A, Gal Y, Cipolla R, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, Salt Lake City, UT, 2018, pp 7482–7491

37. Huo Z, Xia Y, Zhang B (2016) Vehicle type classification and attribute prediction using multi-task RCNN. In: 9th International congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI), Datong, pp 564–569
38. Dvornik N, Shmelkov K, Mairal J, Schmid C. (2017) Blitznet: A real-time deep network for scene understanding. In: Proceedings of the IEEE international conference on computer vision pp 4154–4162
39. Cheng Z, Wang Z, Huang H, Liu Y. (2019) Dense-acssd for end-to-end traffic scenes recognition. In: 2019 IEEE Intelligent Vehicles Symposium (IV) Paris, France, pp 460–465